

潜变量

简体中文版

基础概念 (Foundations) • The Valuation Engineer

Bert Craytor • 2026 年 8 月 2 日

版本 1.0.0

DOI: [10.67330/11n3g168](https://doi.org/10.67330/11n3g168)

简体中文译本，由人工智能辅助翻译并经编辑部审校；英文版为权威版本。

形式定义。 潜变量 (latent variable) 是一项未被观测的属性，它影响观测到的结果，其取值必须由它对可观测变量的作用推断得出，而无法直接测量。

直观阐释。 前五个条目在处理特征向量 $\mathbf{z}$ 时，仿佛它可以被完整测量：建筑面积、地块、景观、状况等等，全都以数据列的形式可得。而在实践中，没有任何特征向量是完整的。有些属性不在数据之中，是因为它们难以测量（室内装修品质、平面布局的流畅度、微区位的朝向细节），是因为标准数据源中根本没有它们（MLS 记录不会记载采光品质、隔音效果，也不会记载近期未取得许可的改造），或者是因为它们尚未被认定为相关（某些社区层面的特征，只有在分析揭示之后才被发现能够预测价格）。

这些未被观测却相关的属性，就是潜变量。它们存在，它们影响价格，但它们不在我们回归的数据列中。它们的作用不会因为未被测量而消失；这些作用会累积到模型的残差里，从已纳入变量的视角看它们像是“噪声”，但从底层数据生成过程的视角看，它们其实是有结构的信号。

潜变量是理想化的特征价格框架与杂乱的估价数据现实之间的桥梁。明确承认它们的存在，是诚实地思考模型不确定性、残差结构以及可观测特征解释力边界的前提。

估价师在何处遇到它。 估价师在三类反复出现的情境中遇到潜变量。

品质与状况调整。 作为估价类别的“品质”与“状况”，本身就是把一个多维潜在概念压缩到单一序数量表上的尝试。UAD 3.6 中的 C3 状况评级，意在概括一宗房地产的递延维修、饰面磨损、机电系统年限与呈现状态——四到十项底层属性被浓缩为一个数字。这个单一数字的概括是潜在概念的代理变量；估价师的功力之一，就在于判断该代理在每个具体案例中对底层真相的捕捉程度。

网格中未获解释的残差。 当三个经调整的可比案例给出价值指示，彼此的差距却超出预期时，这一分歧本身是有信息的。它很少全部是抽样噪声；更常见的情况是，它反映了估价师尚未识别、尚未调整的某项属性。追查这种残余分歧——“为什么这个可比案例的指示值比其他高出 \$50,000？”——正是估价师对潜变量推断的非正式版本。

谁也看不见的近期改造。 一宗房地产可能刚做过厨房翻新、屋顶更换，或建成了经许可的附属住宅单元 (ADU)，而这些尚未反映在 MLS 或税务评估数据中。从估价师的视角看，这项改造实地可见，却不在数据之中。从回归的视角看，这项改造是潜在的。估价师必须把实地观察与数据驱动的指示相互调和，二者都应收敛于该潜在属性的真实贡献。

为何关乎可辩护性。 潜变量带来了若干可辩护性方面的考量，而条目 005 中的可加线性模型并未直接处理这些问题。

遗漏变量偏差。 如果某个潜变量与一个已观测变量相关，它的效应就会渗入该已观测变量的系数。如果近期成交的帕西菲卡 (Pacifica) 住宅往往拥有翻新过的厨房（一项潜在属性），而房屋成交的合同日期又在数据之中，那么成交时点（即 `sale_age`，“距成交天数”）的系数就会吸收一部分厨房翻新效应。`sale_age` 的隐含价格贡献估计值因此被扭曲——它反映的是 `sale_age` 加上未被测量的厨房翻新状态的一部分。可辩护的估价必须承认，已观测变量的系数是以模型中缺失的内容为条件的。

残差的解读。 标准回归理论把残差视为均值为零的高斯噪声。潜变量视角则把残差视为噪声与信号的混合：一部分是抽样变异，另一部分是未观测属性的累积效应。分辨二者在方法论上具有实

质意义。若某残差与位置、与成交日期、与估价师身份、或与任何其他可观测量相关，就提示相应的潜变量可以被部分还原。

模型选择。纳入更多已观测变量是遗漏变量偏差的显见对策，但最要紧的那些潜变量，往往恰恰是最难加入模型的（状况的细颗粒、近期改造的品质、社区微效应）。可辩护的应对方式有时是承认模型永远无法捕捉它们，从而把未获解释的方差归入残差，而不是强行塞进额外的变量。

估价实例演算。条目 005 的回归对可比案例 H 产生了一个异常残差：模型对可比案例 H 价格的低估达 \$46,455，而其余七个残差都落在 $\pm \$18,000$ 的区间内。可比案例 H 的已观测特征——建筑面积 1,600、地块 7,300、无景观、状况“良好”——把它牢牢置于样本的中间位置；其可观测量轮廓中没有任何东西预示异常高的价格。然而它的成交价却比模型所说的高出 \$46,455。

自然的假设是：可比案例 H 具有某项潜在属性，它对价格有贡献，却没有出现在数据中。在本期数据集的构造中，情况恰恰如此：可比案例 H 经历了一次近期厨房翻新，为其组合价值增加约 \$60,000。这项翻新不在回归的变量清单之内，因此其贡献累积在残差中，而不是体现在某个系数上。估计出的残差（\$46,455）略小于真实的潜在贡献（\$60,000），因为回归会把潜在效应的一小部分分摊到其他系数上，在相关维度上使它们向上偏高。

由此可得三点操作性结论：

把离群残差当作证据，而不是误差。对可比案例 H 的残差，第一个恰当的反应不是把它作为离群值剔除，而是去调查这宗房地产的何种特质解释了这一低估。在本例中，实地查勘或建筑许可记录检索就会揭示那次翻新。模型的失败具有诊断价值。

认识到其余系数存在轻微偏差。景观、状况与地块的系数，通过与可比案例 H 已观测特征之间的相关性，吸收了翻新效应的一小部分。在 $n = 8$ 时这一偏差很小，随着可比案例与协变量的增加还会进一步缩小，但它并不为零。

残差整体承载着可还原的结构。这正是作者所提出的残差约束方法（Residual Constraint Approach, RCA）的实质主张（Craytor, 2025）：在住宅估价数据中，残差在系统上并非独立同分布的高斯噪声。它们反映了未观测属性的结构化贡献——状况的细颗粒、品质的细颗粒、功能效用以及其他潜在构念——而这些贡献可以通过把残差本身当作有待建模、而非弃置的潜变量信号来部分还原。

RCA 对状况与品质的处理，使这一两层结构得以落地。估价师的序数评级——即上文讨论的粗颗粒代理变量，可标记为状况-1（状况粗颗粒）与品质-1（品质粗颗粒）——粗糙到几乎可以把它们本身也视为潜变量，但它们仍可作为回归元纳入，以得到近似的隐含价格贡献。随后，该协议把这些估计出的贡献从销售网格回归部分的已测属性中移出、送回残差，再由残差重新分配给细颗粒潜变量状况-2（状况细颗粒）与品质-2（品质细颗粒），并对照 MLS 公开备注等补充证据加以精细化。粗颗粒评级提供了初始的价值分配；细颗粒潜变量则承载可还原的结构。最终，细颗粒的价值贡献取代粗颗粒的价值贡献。

这里的单一可比案例演示，是 RCA 直觉最简单的情形。有了更多数据与更丰富的建模，残差中的结构可以被划分为若干可辨识的潜在因子，每个因子都以标准特征价格设定无法企及的方式对组合的市场价格作出贡献。潜变量不是需要被控制掉的麻烦；它们是可辩护估价建模的下一个前沿。

参考文献。

- Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *American Journal of Psychology*, 15(2), 201–292.
- Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, 53, 605–634.
- Craytor, B. (2025). Residual Constraint Approach: Framework & Protocol. Zenodo. doi:10.5281/zenodo.14787917

交叉引用：特征空间（*characteristics space*）；特征价格函数（*hedonic price function*）；隐含价格（*implicit price*）；特征价格回归（*hedonic regression*）。

附录：Prolog 形式化

以下机器可读的配套文件 (006-latent-variable.pl) 在本刊的 Prolog 知识库中对本条目进行了形式化。该文件同时作为可浏览的 Prolog 版本随本文一并发布。

```
%% 006-latent-variable.pl
%%
%% Prolog companion to Foundations Vol. 1, Issue 1, Article 006
%% Title: Latent Variable
%% Author: Bert Craytor
%% Concept: latent_variable
%% Version: 0.1.0-draft

% =====
% DICTIONARY IMPORT
% =====

:- use_module('.././.././../tools/dictionary/dictionary').
:- dictionary_release('dictionary-2026.1').

% --- From the dictionary (loaded via :- use_module above) -----
% term(latent_variable, "Latent Variable", measurement_theory,
%      "A latent variable is an unobserved attribute that influences
%      observed outcomes, whose value must be inferred from its
%      effects on observable variables rather than measured directly.",
%      published, 'dictionary-2026.1').
%
% term(residual, ...).
% term(omitted_variable_bias, ...).
% term(proxy_variable, ...).
% term(residual_constraint_approach, ...). % synonym: rca
%
% introduced_by(latent_variable, spearman_1904).
% formalized_by(latent_variable, bollen_2002).
% introduced_by(residual_constraint_approach, craytor_2025).
% depends_on(latent_variable, characteristics_space).
% -----

:- use_terms([ latent_variable, residual, omitted_variable_bias,
               proxy_variable, residual_constraint_approach,
               characteristics_space, hedonic_price_function,
               implicit_price, hedonic_regression ]).
:- use_relations([related/2, depends_on/2,
                 introduced_by/2, formalized_by/2]).

% =====
% ARTICLE SUBJECT
% =====

article_concept(latent_variable).
article_issue(volume(1), number(1), year(2026)).
article_id('foundations.2026.006').

% =====
% LOCAL VOCABULARY
% =====

local_term(unobserved).
local_term(noise).
```

```

local_term(signal).
local_term(structured_signal).
local_term(uad_3_6).
local_term(c3_condition_rating).
local_term(mls).
local_term(permit_records).
local_term(neighborhood_micro_effect).

% =====
% THE CENTRAL CLAIM
% =====

claim(latent_variables_accumulate_in_residuals,
      "Unobserved-but-relevant attributes do not vanish when omitted from the model; their effects
      ↪ accumulate in the residual, appearing as noise from the model's perspective but as structured
      ↪ signal from the data-generating process's perspective.").

claim(residuals_are_mixture,
      "Standard regression theory treats residuals as iid Gaussian noise; the latent-variable view
      ↪ treats them as a mixture of sampling noise AND signal from unobserved attributes.").

% =====
% THREE CONTEXTS WHERE APPRAISERS ENCOUNTER LATENT VARIABLES
% =====

appraiser_encounter(quality_and_condition_proxies,
                    "UAD 3.6 'C3 condition' is a single ordinal scale collapsing a multi-dimensional construct (
                    ↪ deferred maintenance, finish wear, mechanical-systems age, presentation). The rating is a
                    ↪ proxy_variable for a latent construct.").

appraiser_encounter(unexplained_grid_disagreement,
                    "When three adjusted comps disagree by more than expected, the disagreement is rarely all
                    ↪ sampling noise; usually reflects an attribute not yet identified or adjusted for.").

appraiser_encounter(unobserved_recent_improvement,
                    "A recent remodel, roof replacement, or permitted ADU buildout may be observable in person but
                    ↪ not in MLS/assessor data — observable to the appraiser but latent to the regression.").

% =====
% DEFENSIBILITY ISSUES
% =====

defensibility_issue(omitted_variable_bias,
                    "If a latent variable is correlated with an observed variable, its effect leaks into the
                    ↪ observed variable's coefficient. Observed-variable coefficients are conditional on what is
                    ↪ missing from the model.").

defensibility_issue(residual_interpretation,
                    "Treat residuals as a mixture, not pure noise. A residual that correlates with location, sale
                    ↪ date, appraiser identity, or any other observable suggests the latent variable can be
                    ↪ partially recovered.").

defensibility_issue(model_selection_limits,
                    "The latent variables that matter most are often hardest to add (condition fine-grain, quality,
                    ↪ micro-locational effects). Sometimes the defensible response is to allocate the unexplained
                    ↪ variance to the residual rather than force-fit additional variables.").

% =====
% WORKED EXAMPLE: COMP H'S ANOMALOUS RESIDUAL

```

```

% =====

:- consult('../..kb/pacifica_comps').

% Comp H's regression residual from entry 005.
comp_h_residual_dollars(46455).

% The latent attribute revealed by investigation.
comp_h_latent_attribute(kitchen_remodel).
comp_h_latent_attribute_value_estimate(60000).

% Why the estimated residual is smaller than the latent attribute's
% true contribution: regression spreads small portions of the latent
% effect across correlated coefficients, biasing them upward.
residual_smaller_than_truth_explanation(
    "Regression spreads small portions of the latent effect across the other coefficients via
    ↪ correlations, biasing them slightly upward in correlated dimensions.").

% =====
% THREE OPERATIONAL CONSEQUENCES
% =====

operational_consequence(outlier_residual_is_evidence,
    "The first response to Comp H's residual is to investigate, not to drop. An in-person
    ↪ inspection or permit search reveals the remodel; the model's failure has diagnostic value.").

operational_consequence(other_coefficients_slightly_biased,
    "View, condition, and lot coefficients absorbed small fractions of the remodel effect via
    ↪ correlations with Comp H's observed characteristics. Bias is small at n=8 but not zero.").

operational_consequence(residuals_carry_recoverable_structure,
    "Residuals in residential appraisal data are systematically not iid Gaussian. They reflect
    ↪ structured contributions of unobserved attributes that can be partially recovered: this is the
    ↪ Residual Constraint Approach intuition.").

% =====
% CONNECTION TO RESIDUAL CONSTRAINT APPROACH (RCA)
% =====
%
% The single-comp illustration here is the simplest case of the RCA
% intuition. With richer modeling, residual structure is partitioned
% into recognizable latent factors.

framework_reference(residual_constraint_approach, craytor_2025).

rca_intuition(
    "Treat the residual itself as a latent-variable signal to be modeled rather than discarded.
    ↪ Latent variables are not a nuisance to be controlled away; they are the next frontier of
    ↪ defensible valuation modeling.").

rca_factor_examples(
    [condition_fine_grain, quality_fine_grain, functional_utility]).

% Two-level condition/quality structure (RCA protocol; mirrors the prose).
rating_levels(condition, condition_coarse_grain, condition_fine_grain).
rating_levels(quality, quality_coarse_grain, quality_fine_grain).

coarse_to_fine_grain_protocol(
    "Regress on the appraiser's coarse-grain C1-C6/Q1-Q6 ratings to obtain approximate value

```

```

↳ contributions; move those contributions out of the measured attributes back into the residual;
↳ reassign them to the fine-grain latent variables and refine against MLS public remarks. In
↳ the end, the fine-grain value contribution replaces the coarse-grain value contribution.

% =====
% INFERENCE RULE: anomalous-residual detection
% =====
%
% A property's residual is "anomalous" if its magnitude exceeds the
% typical residual band reported in entry 005 (+/- $18,000 for the other
% seven comps). Useful for flagging properties whose hedonic prediction
% suggests a latent attribute warrants investigation.

typical_residual_band(18000).

anomalous_residual(P, R) :-
    comp_h_residual_dollars(R0), % stand-in: this entry references the
                                % residual computed in entry 005; future
                                % integration would pull from there.
    P == comp_h,
    R = R0.

% =====
% CROSS-REFERENCES
% =====

cross_reference(characteristics_space).
cross_reference(hedonic_price_function).
cross_reference(implicit_price).
cross_reference(hedonic_regression).

```